Introduction to *Statistical Tests of Equivalence*

Andrew Winterman

The following is an excerpt from my undergraduate thesis. The full text is available, digitally or in print, upon request.

Abstract

Tests of equivalence are motivated and described in detail. Use of tests of equivalence for microarray data analysis is described. A non-parametric bootstrap test and a t-test of equivalence are described and compared against one another via simulation. The bootstrap test, although more computationally expensive, is found to be superior, even in cases when data is normal.

1 Test of Equivalence in Context

Tests of equivalence are a class of statistical hypothesis tests which are designed to assess, through random sampling, whether a parameter of a population's probability distribution is close to a given value. They are not a new statistical idea. Most commonly used by manufacturers of pharmaceuticals, tests of equivalence have found widespread use in the industry since the early 1980s, following a 1979 Food and Drug Admistration decision stating that new, generic[1] drugs would be approved if they could be shown to be "bioequivalent" to existing, approved drugs [15]. As a matter of course, both the generic and patented original version of the drug are nominally chemically identical, but they might differ in a number of ways, including milling procedure, delivery system, etc. Hence it must be shown that a new generic acts upon the body in ways sufficiently similar to the patented original – that is, that the new drug is bioequivalent to the old one [4]. The FDA mandated that bioequivalence be demonstrated by clinical trials employing tests of equivalence. Because of the billions of dollars involved in the generic drug industry, there has been substantial theoretical attention, and many biopharmaceutical applications thoroughly described (for an exposition of such applications, see Wellek [15]).

The purpose of this thesis is to describe two such tests of equivalence, and extend their implementation to the context of DNA microarray experiments, a

[1] Once a drug has exhausted the duration of its patent, it can be reproduced by any company, often at a fraction of the price of the patented version. These reproductions are called "generic" drugs.

scientific tool in the field of genetics. In this application, tests of equivalence have not been in common use - a Web of Science search (Summer 2010) has uncovered only two papers addressing the subject: [9] and [6] - but are necessary to the corroboration of common and desirable claims.

2 The Two Color DNA Microarray

Microarrays are designed to compare the amounts of gene-coding fragments of DNA or RNA found in different tissues - that is, relative gene expression between the two tissues - through a phenomenon known as DNA hybridization. Essentially, hybridization refers to a process in which single strands of complementary DNA or RNA chemically bind to one another. Non-complementary strands sometimes bind together as well, but a better fit induces more fragments to bind.

To take advantage of hybridization, microarrays, small glass slides, are imprinted with thousands of fragments of DNA. Many copies of each fragment are placed in high density on a clearly delineated spot, which is referred to interchangeably as a feature or a probe. Ideally, the DNA in each probe contains the complete code for a gene of interest, although this may not always be the case. A gene is defined as a segment of DNA which codes for the production of a particular protein. Apart from the technical difficulty of slicing up DNA at precise points, it can also be difficult to tell exactly where one gene starts and another ends. The features on the array may or may not contain the full genetic code for a given gene.[2] Despite such considerations, the DNA fragment imprinted in a feature is called a gene.

A two color microarray experiment can be divided into two steps. First the complement of the DNA (or RNA) which encodes the genes of interest to the scientist is imprinted on the microarray itself. Then, DNA is taken from two different sources, dyed two colors, and then washed over the prepared microarrays. The dyed DNA is induced to hybridize, i.e. bind to the DNA in the probes on the array. Then, lasers are used to assess relative intensities of fluorescence of the two dyes coloring the features on the array.

For example, if we were comparing gene expression across two phenotypes[3] of a lemur species, say lemur phenotype A and lemur phenotype B, we would prepare the slide with fragments of DNA which we suspect to be present in both phenotypes. Then we would extract DNA from A and B and dye the DNA red (Cyanine 5 dye) and green (Cyanine 3 dye) respectively[4]. We would wash our solutions of DNA and dye over the microarrays, and induce the DNA to hybridize. Finally we would shine a green laser and a red laser at the features on the microarray slide. The intensity of the reflected red or green light from

[2] See [16] for an algorithm which attempts to solve this problem.

[3] A phenotype is an observable trait of an organism, such as morphology or behavior. In this context the the word means a group of organisms exhibiting the observable trait.

[4] Actually, Cyanine 5 is simply a little more reddish and Cyanine 3 is a little more greenish. The laser used to assess intensity is, in any case, precisely calibrated to the frequency of light each dye reflects.

each feature is proportional to the amount of DNA of that color bound to the feature. Hence the ratio of reflected light from a given feature is the ratio of the amount of DNA (which matches the DNA in the feature) from each lemur in the solutions washed over the slide [13]. If done properly, the ratios of reflected light is linearly related to the relative expression of the given DNA fragment in each of our lemurs.

Each feature on a microarray slide produces a single data point in a single experiment. Hence if a slide has six thousand features, running it will produce a single data point in each of six thousand experiments (unless of course, genes are repeated on each slide). The experimenter must use quite a few slides to obtain a suitable number of data points in each experiment. As microarrays can be expensive, non-commercial microarray experiments often involve many relatively small data sets. They remain computationally intensive - often requiring the analysis of several thousand experiments at once - meaning development of effective tools for analysis of these data sets requires some mathematical training.

The statistical tools commonly employed in the analysis of microarray data are hypothesis tests of difference, such as the two sided t-test, which assumes there is no difference between the two data sets and looks for contradictory evidence. If no evidence of difference is found - that is, if a significance level α test finds that the given gene is *not* differentially expressed - researchers are often tempted to conclude that the gene is equivalently expressed across the tissues under examination at level α. However, the assertion that genes not significantly different are significantly equivalent is a logical fallacy born of a seductive misunderstanding of hypothesis testing. A hypothesis test requires the specification of a null hypothesis - that is, a set of non-contradictory assumptions - against which we hope to find evidence. In the microarray context, the null hypothesis might be that there is no difference in gene expression levels across the two groups. We can then either reject or fail to reject this null hypothesis. If we fail to reject the null, we do not have sufficient reason to conclude that the null is true. We have no evidence to the contrary, but we also do not necessarily have any in support[5] [2]. Hence a test of difference, which takes as its null hypothesis the assumption of no difference, does not necessarily supply evidence of equivalence.

In the absence of a fully developed test of equivalence, researches are left with ad hoc methods to determine equivalence, which often involve the logical error described in the preceding paragraph. For example, in Adjaye et al. (2004) [1], the authors describe techniques for assessing Bovine gene expression using Microarrays prepared with stretches of human DNA. The goal of their paper is to demonstrate that certain fragments of Human and Bovine DNA bind equally well to Microarrays prepared with human DNA. After running three different statistical tests to assess significance of differential expression, Adjaye et al. assert, "thus, we conclude that the level of expression of the in-

[5]It is possible to design a test which, when the probability of type II error is sufficiently low, can find evidence for the null. However, in general this is a reckless approach, and logically incorrect besides.

dividual 349 genes under investigation within human and bovine brain [sic] is roughly the same." Price et al (2008) commit a similar error in search of genes whose expression during the aging of the Wallower (Erysimum linifolium) [8] remains unchanged between leaves and flowers. They apply a Student's t test of difference and find, "for 263 probes derived from the leaf cDNA library, 52% showed up regulated expression with age in leaves, although larger numbers of leaf-derived probes on the array were stable in expression with leaf senescence." Finally, Rodriguez-Lannety et al (2007) [12] provide guidance as to how one might select housekeeping genes - endogenous experimental controls whose expression remains unchanged across sources - for normalization of microarray experiments in the field of coral and cnidarian biology. The authors "tested whether the Cy5/Cy3 ratios were not significantly different to ratio = 1 (null hypothesis) using a one sample t-test ... filtering out those genes whose ratios were significantly different from one ($p < 0.05$)." That is, they applied a test of difference as a filter, and asserted that those genes not significantly different are genes which might be equivalently expressed[6]. They go on to apply several other methods to ascertain which genes have similar expression levels.

All the authors cited above would like to apply a test of equivalence. Without this tool it is difficult to correctly compute the p-values for their results, which hinders the comparison of results from multiple experiments. As both equivalent and differential expression of genes are of interest, tests of equivalence should be a part of any microarray analyst's statistical tool kit. This thesis contains a careful exposition of two different methods for carrying out a test of equivalence, including explicit application to microarray data analysis.

3 Two Tests in Broad Brush

Let us suppose we are examining gene expression between the lemur phenotypes A and B as above. A statistician asked to assess whether or not A and B are equivalent would first choose a parameter, Δ, which measures the difference between the two groups, and for that parameter an estimator, D, whose value, should gene expression in A be the same as gene expression in B, is known. She calls the value of Δ given no difference ζ. She would employ two real numbers, $\epsilon_1 < \zeta$ and $\epsilon_2 > \zeta$ to specify how similar gene expression in the two species must be to one another to be considered *equivalent* for her purposes. In other words, if it is true that:

$$\Delta \in (\epsilon_1, \epsilon_2),$$

then we say A and B are equivalent. A test of equivalence assesses whether or not the difference as measured by Δ is substantive – two samples are declared equivalent if and only if the difference between them is so small so as to not matter.[7]

[6] This careful proposition is correct, for the appropriate test of equivalence

[7] It is still left to the scientist to determine what constitutes a difference that matters.

To this end, one might compute D and an associated confidence interval, I, of appropriate level; rejecting the hypothesis of non-equivalence should that confidence interval lie completely inside some critical interval, i.e. (ϵ_1, ϵ_2).

In Chapter 2, a two one sided Student's t-test (a parametric test) and a bootstrap test (a non-parametric test) are described. The t-test assumes normally distributed measurement error, an assumption not necessarily well founded. The bootstrap relies on fewer assumptions but is more computationally intensive. In Chapter 3 the implementation of both tests in the microarray context is described. In Chapter 4 the power of both tests is assessed using artificial data sets. Chapter 5 concludes by suggesting future areas of research.

References

[1] J. Adjaye, R. Herwig, D. Herrmann, W. Wruck, A. Benkahla, T. C. Brink, M. Nowak, J. W. Carnwath, C. Hultschig, H. Niemann, and H. Lehrach. Cross species hybridization of human and bovine orthologous genes on high density cdna microarrays. *BMC Genomics*, 5:83, 2004.

[2] D.G. Altman and J.M. Bland. Absence of evidence is not evidence of absence. *British Medical Journal*, 311:485, 1995.
Yoah Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[3] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Inc., Oakland, California, first edition, 1997.

[4] D.J. Birkett. Generics - equal or not? *Australian Prescriber*, 26(4), 2003.

[5] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, 1993.

[6] W. J. Eijgelaar, A. J. Horrevoets, A. P. Bijnes, M.J. Daemen, and W. F. Verhaegh. Equivalence testing in microarray analysis: similarities in the transcriptome of human atherosclerotic and nonatherosclerotic macrophages. *Physiological Genomics*, 41:212 –223, January 12 2010.
Stephen H. Friedberg, Arnold J. Insel, and Lawrence E. Spence. *Linear Algebra*. Pearson Education, Inc., Upper Saddle River, New Jersey 07458, fourth edition, 2003.

[7] Ronald R. Hocking. *The Analysis of Linear Models*. Wadsworth Inc., Belmont, California 94002, first edition, 1985.

[8] A. M. Price, D. F. Aros Orellana, F. M. Salleh, R. Stevens, R. Acock, V. Buchanan-Wollaston, A. D. Stead, and H. J. Rogers. A comparison of leaf and petal senescence in wallflower reveals common and distinct patterns of gene expression and physiology. *Plant Physiology*, 147:1898 – 912, 2008.

[9] Jing Qiu and Xiangqin Cui. Evaluation of a statistical equivalence test applied to microarray data. *Journal of Biopharmaceutical Statistics*, 20(2):240 –266, March 2010.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[11] Susan C. P. Renn, Nadia Aubin Horth, and Hans A Hofmann. Fish and chip: functional genomics of social plasticity in an African cichlid fish. *The Journal of Experimental Biology*, 211:3041–3056, 2008.

[12] M. Rodriguez-Lannety, W. S. Phillips, S. Dove, O. Hoegh-Guldberg, and V. M. Weis. Analytical approach for selecting normalizing genes from a cdna microarray platform to be used in q-rt-pcr assays: a cnidarian case study. *Journal of Biochemical and Biophysical Methods*, 70:985–991, 2007.

[13] Dari Shalon, Stephen J. Smith, and Patrick O. Brown. A DNA microarray system for analyzing complex DNA samples using two - color fluorescent probe hybridization. *Genome Research*, 6:639–645, 1996.

[14] Gordon K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. Huber R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

[15] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2010.

[16] Anthony Ian Wirth. *A Plasmodium falciparum Genefinder: Honours research project*. Honours thesis, University of Melbourne, 2000.